Isolated word recognition using neural network

Conference Paper · December 2015

DOI: 10.1109/INDICON.2015.7443697

4 authors, including:

Sarfaraz Masood
Jamia Millia Islamia
31 PUBLICATIONS **118** CITATIONS
SEE PROFILE

Danish Raza Rizvi
Jamia Millia Islamia
6 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Isolated Word Recognition Using Neural Network

Sarfaraz Masood, Danish Raza Rizvi
Department of Computer Engineering
JamiaMilliaIslamia
New Delhi,India
sarfarazmasood2002@yahoo.com
drizvi@jmi.ac.in

Madhav Mehta,Namrata
Department of Computer Engineering
JamiaMilliaIslamia
New Delhi, India
madhavmehta16@gmail.com
namrata.srs@gmail.com

***Abstract*: Isolated Word Recognition is the process of converting the spoken word into its corresponding text format. At present mainly Mel Frequency Cepstrum Coefficients (MFCC) is used as the feature extraction parameter i.e. the identifying features for the speech signal. Through this paper efforts have been made to determine the accuracy of an MFCC based system and also to build an isolated word recognizer based on word acoustic model that uses MFCC in combination with other features of speech such as Root Mean Square Energy, Length of the word and its Brightness. Using an artificial neural network as the classifier, the system was trained & tested for a set of spoken isolated words. The results obtained showed a high and an increased accuracy for the experiment in which along with MFCC other selected parameters were also involved against the experiment which only involved MFCC.**

Keywords-Word Recognition; Mel Frequency Cepstrum Coefficient; word acoustic model; root mean square energy; length; brightness

I. INTRODUCTION

Technology has become an integral part of our lives. New developments continue to take place in the field of information and technology. Scientists worldwide are making efforts to emulate human intelligence in machines. Word and speech recognition is one such method that allows machines to exhibit human intelligence. Implementation of techniques such as Word Recognition would be a step towards revolutionizing the spread of IT industry. Also applications based on this technique would prove to be a boon for people with visual impairment for whom keyboard is an obstacle in accessing the computer.

In isolated word recognition system the input is taken as individual words or a list of words separated by a defined pause. Each word is processed individually and the recognition of a word does not affect and is not affected by words preceding or succeeding it. While a continuous speech recognizer takes a continuous input of speech and the recognition of words may require contextual understanding of similar sounding words.

The isolated word recognizer involves taking a word input in the form of a wav file. The wav file is then processed into a series of acoustic vectors. The word recognizer requires a recognizing unit which may be a neural network or a hidden markov model that takes the acoustic vector as an input and maps it into the corresponding word.

A lot of work has been done in the field of word recognition but substantial amount of work still needs to be done. This has motivated the authors of this paper to build a word recognizer. Also most of the work in this field of speech and word recognition involves the use of MFCC as the identifying features. So through this paper an attempt is made to determine the accuracy of an MFCC based recognition system and to check if this accuracy can further be improved by using other parameters of speech along with MFCCs.

The paper is organized as follows: Section 2 discuss the recent studies done in related work done in this field. Section 3 describes the proposed system design for the isolated word recognizer. This is followed by section 4 which describes the features extracted in the dataset for the purpose of constructing a recognizer. In section 5, the experimental design is explained and finally in section 6 the results obtained from the experiments are discussed which is followed by the conclusion.

II. RELATED WORK

The field of speech and word recognition has invoked the interest of many people and a lot of noteworthy work has been done in this area. Some of the promising works are included in this section.

Preeti Saini, Parneet Kaur, Mohit Dua have developed an automatic Hindi speech recognition system using HTK [1]. The system development was done by using Hidden Markov Model Toolkit (HTK). MFCC are used as the identifying feature of speech. This System is trained and can recognize 113 words only. The system works well in both speaker dependent and speaker independent context. An application specific continuous speech recognition system has been developed in Hindi by Gaurav Devanesamonim, Shakina Devi, Gopal Krishna Sharma and Mahua Bhattacharya [2]. It is a speaker independent phone based system that uses MFCC as the characteristic speech parameter and Hidden Markov Model for feature extraction and development of the model. It uses phonemes to map words. The system involves usage of 29 phonemes at present. Other features apart from MFCC have also been used for developing recognition system. Tarun Pruthi, Sameer Sakena and Pradip K Das have implemented isolated word recognition system in Hindi using Vector Quantization (VQ) and Hidden Markov Model [3]. Linear Predictive Coding (LPC) and Vector Quantization are done for feature extraction and then recognition is done using Hidden Markov Model. It is a speaker dependent

system that uses recording from two male speakers. The vocabulary consists of Hindi digits from 0 to 9 and for each of these digits 20 utterances are recorded for the two speakers.

Apart from these works many popular commercial speech recognition application have been developed such as Dragon Naturally Speaking, IBM Via voice and Microsoft SAPI.

Also many comparative studies have been done which check the performance of MFCC technique against other techniques. H. B. Chauhan, Prof. B. A. Tanawala in their paper Comparative Study of MFCC and LPC Algorithms for Gujrati Isolated Word Recognition have compared the performance of MFCC with Linear Predictive Coding (LPC) features under vector quantization [4]. The system was trained for 200 utterances of the words and the system accuracy was calculated for MFCC and then for LPC. The MFCCs proved to be a more accurate technique.

III. PROPOSED SYSTEM DESIGN

The word recognition system consists of the following components:

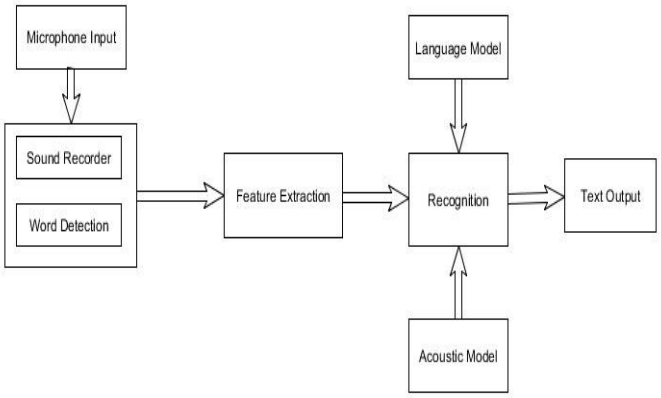

Figure 1: Block Diagram of Recognition System [5]

A. Input

A microphone is used to record the spoken word in a closed room. Sampling is done at a rate of 16000 Hz. The word or the speech is recorded in the form of a wav file. This file is then filtered to remove any kind of distortion or noise that may be present and then it is passed to the feature extraction component.

B. Feature Extraction

Human ears can differentiate one word from another word. This is because each word has certain characteristic features which help to recognize the word. In this paper the various features that are extracted from a word are- Mel Frequency Cepstrum Coefficient, Root Mean Square Energy (RMS) or the energy associated with the word, Brightness and the Length of the word measured in seconds.

C. Recognition

The recognition component is the heart of the system. The various features that are extracted from a word are fed to the recognizing component. For this paper a supervised neural network is used as the recognizing unit.

D. *Language Model*

The language model is used in phone based speech recognition systems to identify similar sounding phones. It allows formation of valid sequences by considering the context. In this paper a word based acoustic model was used so there was no need to implement a language model.

E. *Acoustic Model*

The paper is based on a word based acoustic model i.e. the system is capable of recognizing only those words for which it has been trained. The word acoustic model is used for implementing a small vocabulary as against the phone model that is used for implementing a large vocabulary by modeling the various phone and the whole word is written as a chain of phones.

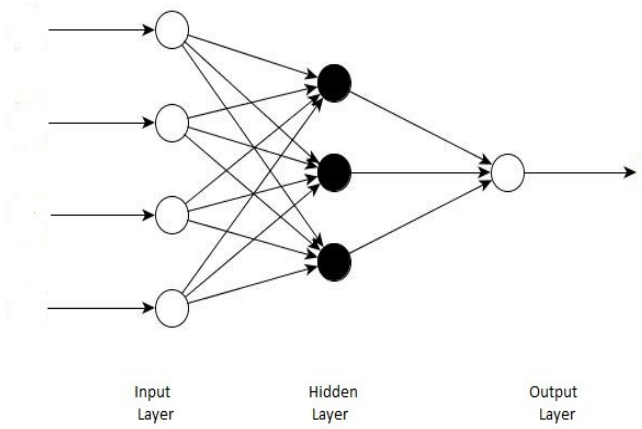

Figure 2: Diagram of a multilayer feed forward network [9]

For the purpose of constructing a classifier, a multilayer feed forward artificial neural network was trained and tested upon. A multilayer feed forward network is a network consisting of a number of interconnected layers and some weight is associated with each connection. Any layer apart from input and output layer is called as the hidden layer as it is internal to the system and has no interaction with the outside world. The input layer and output layer are involved in receiving the input and computing the output. A feed forward neural network is one in which the output of a node is an input to the succeeding node only and is not reverted to the preceding or the same node [10].

IV. FEATURE EXTRACTION

A. Mel Frequency Cepstrum Coefficients

The same word even when spoken by different speakers is perceived to be the same by humans because of the distinctive characteristics that categorize a word. Of all the features known to us Mel Frequency Cepstrum Coefficients resemble the human ear the most and thus are most frequently used. The MFCC are used to convert the speech signal into a set of discrete acoustic vectors that can be used to characterize the word input.

B. Selecting Other Parameters

For the purpose of this paper to compare the accuracy of an MFCC based system with a system using MFCC together with some other parameters a set of parameters had to be chosen. A number of parameters of speech can be calculated and it was not possible to determine the accuracy of system for each and every one of these parameters. So intuitively a few parameters of speech were selected. Those parameters were chosen which appeared to have a direct relation with the word utterance such as energy of the word, its length and its brightness.

A. Root Mean Square Energy

It measures the energy of the audio file or the spoken word by evaluating the root average of the square of the amplitude Different words have different energy or intensity associated with them and thus it can be an identifying characteristic for words.
The Root Mean Square Energy is given by the following equation.

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} xi^2} = \sqrt{\frac{x1^2 + x2^2 + \ldots\ldots + xn^2}{n}} \qquad (1)$$

B. Length

The Length of a word is a measure of its length in seconds. Each word takes a characteristic amount of time to be spoken even when spoken by different people. For instance the two words – Kiwi and Pineapple have different amount of time length. And the relative difference in the length of these words even when spoken by different people remains same.

C. Brightness

The brightness of a word measures the high frequency energy in a word that is the energy above a threshold frequency. It is expressed as a number between 0 and 1.

V. EXPERIMENTAL DESIGN

A recognition system for a vocabulary of seven words is implemented. For each word twenty nine utterances are recorded resulting in 29*7=203 files. Voices of three different people (two male and one female) are used to train the system.

The voice samples for the first speaker (male) were taken from [8]. The voice sample consists of 15 utterances for each of the 7 words resulting in 15*7=105 files. The rest of the voice samples were recorded from two speakers (one male and one female). The recording was done at a sampling rate of 16 kHz in a closed room. For each of the seven words seven utterances from each of the two speakers were recorded. By obtaining the recordings from different speakers the system was trained and tested for a variety of voice samples.

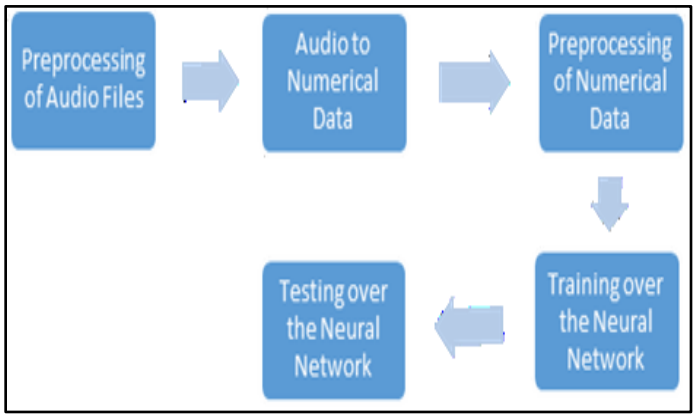

Figure 3: Proposed flow of the experiment

Preprocessing is done so as to remove any kind of noise or distortion from the recorded files. All the four features namely MFCC, root mean square energy, brightness and length are calculated for each file i.e. features in the form of the numerical data are extracted from each audio file

For the purpose of training, 60% of the dataset was used and the rest 40% of dataset was assigned for the purpose of testing the trained system. Normalization was done to both testing and training data so as to obtain the values of features in the range [-1, 1] so that the neural network can produce efficient results. The following table describes the values of the various network parameters of the multilayer perceptron chosen.

Table.1 Network Parameters

Parameter	Value
Number of Input Nodes	07
Number of Hidden Layer	01
Number of Hidden Nodes	10
Training Algorithm	Gradient Descent
Activation Function	Tansigmoidal
Learning rate	0.17
Epochs	1000

Two separate experiments were carried out for the purpose of this work. In order to compare the accuracy of an MFCC based system with a system using all four parameters, first only Mel frequency cepstral coefficients are fed to the neural network as input and accuracy calculated. In the other experiment all the four parameters are fed as an input to the neural network and the accuracy of system on test data is determined. For the purpose of comparison the learning rate, algorithm used for weight updation, number of epochs and the number of hidden layers is kept same in both the experiments.

Any other combination of these parameters may be used for developing the system. But the best result in our case was

obtained for these values.

VI. RESULT

The results were separately evaluated for both the experiments i.e. an MFCC based recognition system and for a system using the four parameters- MFCC, root mean square energy, length and brightness.

For the purpose of comparison values like true positive (TP), true negative (TN), false positive (FP) and false negative (FN) [6] were evaluated. These were further used to calculateaccuracy for each of the seven words in the dataset–Apple, Banana, Kiwi, Lime, Orange, Peach and Pineapple using the values of. Accuracy is given by

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (2)$$

For the onlyMFCC based experiment the following values were observed.

Table.2 The observed values for only MFCC based system

Result	Apple	Banana	Kiwi	Lime	Orange	Peach	Pineapple
FN	1	1	4	5	2	0	0
TN	71	72	72	72	72	72	60
TP	11	11	8	7	10	12	12
FP	1	0	0	0	0	0	12
Acc(%)	97.62	98.81	95.24	94.05	97.62	100	85.71

The overall accuracy for an MFCC based system was calculated by taking the average for all seven words and it is found to be 95.57%.

For the second experiment, which involved MFCC and other extracted features, the following results were obtained

Table.3The observed values for a system using all four parameters

Result	Apple	Banana	Kiwi	Lime	Orange	Peach	Pineapple
FN	0	0	0	7	0	1	0
TN	72	72	72	72	72	72	64
TP	12	12	12	5	12	11	12
FP	0	0	0	0	0	0	8
Acc %	100	100	100	91.67	100	98.81	90.48

As per the results described in Table 3, the accuracy of recognizing the isolated words in our dataset, increased significantly. And for some words like apple, banana, kiwi and orange the accuracy went up to 100%. However the overall accuracy for this system is obtained by taking the average of the accuracy for all seven words and was observed to be 97.28%.

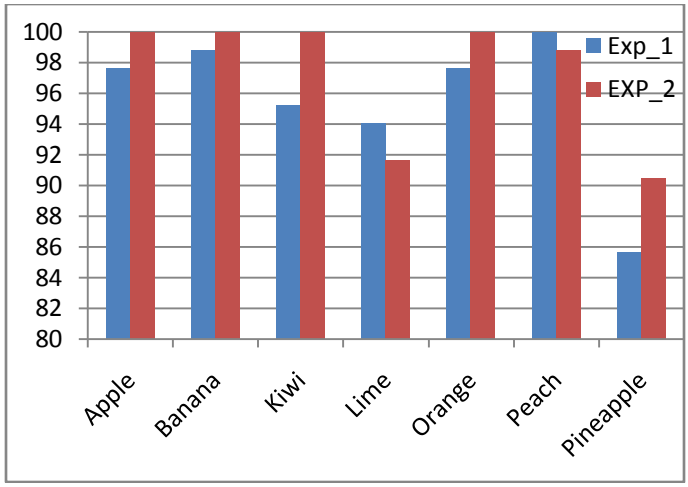

Figure 4: Comparison of Accuracy for both experiments

Hence the accuracy for a system using all the features was observed to be higher than a system using MFCC alone.

VII. CONCLUSION

This work attempted to solve the isolated word recognition problem. Separate experiments were conducted to test only MFCC based word recognition system against the proposed systems that also incorporated some extra acoustic features for the purpose of recognition. The results obtained from the experiments conducted strongly suggest that when MFCC was used in the recognizer along with these proposed acoustic features, the accuracy of the systems was affected significantly. In fact the recognition accuracy of most of the words considered for this work showed a noteworthy rise. Such a system fully developed would go a long way in the spread of IT in the lives of the common people especially those who are constrained by visual disabilities.

Hence MFCC is definitely a very reliable feature for word recognition and efficient systems can be developed by using it. However the efficiency of these systems can be further improved by using the proposed features in addition to MFCC.

RFERENCES

[1] PreetiSaini, ParneetKaur, MohitDua, "Hindi Automatic Speech Recognition Using HTK," International Journal of Engineering Trends and Technology (IJETT) – Volume4 Issue6- June 2013.

[2] Gaurav, DevanesamoniShakina Devi, Gopal Krishna Sharma, Mahua Bhattacharya, "Development of Application Specific Continuous Speech Recognition System in Hindi", Journal of Signal and Information Processing, 2012, 3, 394-401.

[3] TarunPruthi , Sameer Saksena and Pradip K Das, "Isolated Word Recognition for Hindi Language using VQ and HMM" International Conference on Multimedia Processing and Systems (ICMPS-2000), IIT Madras, Chennai, 13-15 August, 2000

[4] H. B. Chauhan, Prof. B. A. Tanawala "Comparative Study of MFCC And LPC Algorithms for Gujrati Isolated Word Recognition", International Journal of Innovative Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2015

[5] Ripul Gupta, “Speech Recognition for Hindi,”, IIT Bombay M.Tech Thesis (2011)
[6] Data Mining: Concepts and Techniques, 2nd ed. Jiawei Han and MichelineKamber
[7] MIRtoolbox Documentation.
[8] Sandsmark, Håkon. "Isolated-word speech recognition using hidden Markov models." (2010).
[9] ShoukatUllah,,ZakiaHussain, “Improve An Efficiency Of Feed Forward Multilayer Perceptrons By Serial Training”, Journal of Theoretical and Applied Information Technology. Vol6. No1. (pp 017-020)
[10] Haykin, Simon, and Neural Network. "A comprehensive foundation." *Neural Networks* 2.2004 (2004).